Sumca: Simple, Unified, Monte-Carlo Assisted Approach to Second-order Unbiased MSPE Estimation

JIMING JIANG AND MAHMOUD TORABI
University of California, Davis and University of Manitoba

We propose a simple, unified, Monte-Carlo assisted (Sumca) approach to second-order unbiased estimation of mean squared prediction error (MSPE) of a small area predictor. The proposed MSPE estimator is easy to derive, has a simple expression, and applies to a broad range of predictors that include the traditional empirical best linear unbiased predictor (EBLUP), empirical best predictor (EBP), and post model selection EBLUP and EBP as special cases. Furthermore, the leading term of the proposed MSPE estimator is guaranteed positive; the lower-order term corresponds to a bias correction, which can be evaluated via a Monte-Carlo method. The computational burden for the Monte-Carlo evaluation is much lesser, compared to other Monte-Carlo based methods that have been used for producing second-order unbiased MSPE estimators, such as double bootstrap and Monte-Carlo jackknife. The Sumca estimator also has a nice stability feature. Theoretical and empirical results demonstrate properties and advantages of the Sumca estimator.

Key Words. approximation, bias correction, Monte Carlo, MSPE, second-order unbiasedness, small area estimation

1 Introduction

In recent years there has been substantial interest in small area estimation (SAE) that is largely driven by practical demands. In policymaking regarding allocation of resources to subgroups (small areas), or determination of subgroups with specific characteristics (e.g., in health and medical studies) in a population, it is desirable that the decisions be made based on reliable estimates. However, information or data collected from each subgroup

are often inadequate to achieve an acceptable degree of accuracy using the direct survey method, which computes the estimate solely based on data from the subgroup. For this reason, demands and interest in research on SAE have kept increasing. Recent reviews on SAE and related topics can be found, for example, in Pfeffermann (2013), Rao and Molina (2015), and Jiang (2017, ch. 4).

A major topic in SAE is estimation of mean squared prediction errors (MSPEs) for predictors of various characteristics of interest associated with the small areas. A standard practice for the MSPE estimation in the current SAE literature is to produce a second-order unbiased MSPE estimator, that is, the order of bias of the MSPE estimator is $o(m^{-1})$, where m is the total number of small areas from which data are collected (see Section 6 for discussion regarding other aspects of the MSPE estimation). For the most part, there have been two approaches for producing a second-order unbiased MSPE estimator. The first is the Prasad-Rao linearization method (Prasad and Rao 1990). The approach uses Taylor series expansion to obtain a second-order approximation to the MSPE, then corrects the bias, again to the second-order, to produce an MSPE estimator whose bias is $o(m^{-1})$. Various extensions of the Prasad-Rao method have been developed; see, for example, Datta and Lahiri (2000), Jiang and Lahiri (2001), Das, Jiang and Rao (2004), and Datta, Rao and Smith (2005). Although the method often leads to an analytic expression of the MSPE estimator, the derivation is tedious, and the final expression is likely to be complicated. The linearization method also does not apply to situations where a non-differentiable operation is involved in obtaining the predictor, such as model selection (e.g., Jiang, Lahiri and Nguyen 2018) and shrinkage estimation (e.g., Tibshirani 1996).

The second approach to second-order unbiased MSPE estimation is resampling methods. Jiang, Lahiri and Wan (2002; hereafter, JLW) proposed a jackknife method to estimate the MSPE of an empirical best predictor (EBP) of a mixed effect associated with the small areas. The method avoids tedious derivations, and is "one formula for all". On the other

hand, there are certain restrictions on the class of predictors to which JLW applies. Namely, for the JLW to be applicable to the empirical best linear unbiased predictor (EBLUP), widely used in SAE, a posterior linearity condition needs to be satisfied. This is not the case, for example, if the random effects involved in a linear mixed model (LMM) for SAE are not normally distributed. Also, JLW does not apply to a predictor that is obtained post model selection. In order to solve the latter problem, Jiang *et al.* (2018) proposed a Monte-Carlo jackknife method (McJack). The method leads to a second-order unbiased MSPE estimator in situations like post-model selection (PMS); however, it is computationally intensive. Another resampling-based approach is double bootstrapping (DB; Hall and Maiti 2006a,b). Although DB is capable of producing a second-order unbiased MSPE estimator, it is, perhaps, computationally even more intensive than the McJack.

The method to be proposed in this paper may be viewed as a hybrid of the linearization and resampling methods, by combining the best part of each method. We use a simple, analytic approach to obtain the leading term of our MSPE estimator, and a Monte-Carlo method to take care of a remaining, lower-order term. The computational cost for the Monte-Carlo part is much less compared to McJack and DB in order to achieve second-order unbiasedness. More importantly, the method provides a unified and conceptually easy solution to a hard problem, that is, obtaining a second-order unbiased MSPE estimator for a possibly complicated predictor.

A heuristic derivation of the MSPE estimator, called Sumca (which is an abbreviation of simple, unified, Monte-Carlo assisted) estimator, is given in the next section, which is followed by rigorous theoretical justification. In Section 3, we consider the special case of PMS predictors. In Section 4, we illustrate the Sumca method with a few examples. In Section 5, we study performance of the Sumca estimator, and compare it with popular existing estimators via Monte-Carlo simulation. A concluding remark is offered in Section 6. Technical and additional empirical results, including real-data applications, are provided

in the Supplementary Material.

2 Derivation and justification

We begin with a heuristic derivation of the proposed MSPE estimator, followed by rigorous justification. The essential idea of the derivation is the same as that used in the seminal paper of Prasad and Rao (1990) and subsequent work on linearization methods as well as resampling methods for higher order unbiasedness (e.g., Jiang *et al.* 2002, Hall and Maiti 2006a). Namely, for a sufficiently smooth function $g(\psi)$ and under regularity conditions, the bias of $g(\hat{\psi})$ for estimating $g(\psi)$ is $O(m^{-1})$ for a root-n consistent (and asymptotically normal) estimator $\hat{\psi}$. If one writes the bias as $c(\psi)/m + o(m^{-1})$, where $c(\psi)$ is another smooth function, one can, again, approximate $c(\psi)$ by $c(\hat{\psi})$ to obtain a bias-corrected estimator whose bias is $o(m^{-1})$. The latter type of bias expansion can also be found, for example, in the context of jackknife (e.g., Shao and Tu 1995, p. 5).

Let θ denote a mixed effect, that is, a (possibly nonlinear) function of fixed and random effects that are associated with a mixed effects model (e.g., Jiang 2007). Let $\hat{\theta}$ be a predictor of θ, that is, a function of the observed data, y. Here, $\hat{\theta}$ may be a traditional predictor like EBLUP, or EBP, or something more complicated like the PMS EBLUP or EBP, to which the standard methods such as Prasad-Rao and JLW do not apply to obtain a second-order unbiased MSPE estimator. The MSPE of $\hat{\theta}$ can be expressed as

$$\mathrm{MSPE} = \mathrm{E}(\hat{\theta} - \theta)^2 = \mathrm{E}\left[\mathrm{E}\{(\hat{\theta} - \theta)^2|y\}\right]. \tag{1}$$

Suppose that the underlying distribution of y depends on a vector of unknown parameters, ψ. Then, the conditional expectation inside the expectation on the right side of (1) is a function of y and ψ, which can be written as

$$a(y, \psi) = \mathrm{E}\{(\hat{\theta} - \theta)^2|y\} = \hat{\theta}^2 - 2\hat{\theta}h_1(y, \psi) + h_2(y, \psi), \tag{2}$$

where $h_j(y,\psi) = \mathrm{E}(\theta^j|y), j = 1, 2$. If we replace the ψ in (2) by $\hat{\psi}$, a root-n consistent and asymptotically normal estimator of ψ, then under regularity conditions such as differentiability and dominatedness (e.g., Jiang 2010, p. 32), the result is a first-order unbiased estimator of the MSPE, that is,

$$\mathrm{E}\{a(y,\hat{\psi}) - a(y,\psi)\} \quad = \quad O(m^{-1}). \tag{3}$$

On the other hand, both $\mathrm{MSPE} = \mathrm{E}\{a(y,\psi)\}$ [by (1), (2)] and $\mathrm{E}\{a(y,\hat{\psi})\}$ are functions of ψ, denoted by $b(\psi)$ and $c(\psi)$, respectively. By (3), we have $d(\psi) = b(\psi) - c(\psi) = O(m^{-1})$; thus, if we replace, again, ψ by $\hat{\psi}$ in $d(\psi)$, the difference would be a lower-order term,

$$d(\hat{\psi}) - d(\psi) \quad = \quad o_{\mathrm{P}}(m^{-1}). \tag{4}$$

Now consider the estimator

$$\widehat{\mathrm{MSPE}} \quad = \quad a(y,\hat{\psi}) + b(\hat{\psi}) - c(\hat{\psi}). \tag{5}$$

One would have, by (1)–(5), $\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{E}\{a(y,\psi)\} + \mathrm{E}\{a(y,\hat{\psi}) - a(y,\psi)\} + \mathrm{E}\{d(\hat{\psi})\} = \mathrm{MSPE} + \mathrm{E}\{d(\hat{\psi}) - d(\psi)\} = \mathrm{MSPE} + o(m^{-1})$. Essentially, this one-line, heuristic derivation shows the second-order unbiasedness of the proposed MSPE estimator, (5), provided that it can be justified and the terms involved can be evaluated (see below).

It should be noted that, strictly speaking, virtually all the subjects we deal with, such as $a(y,\psi), d(\psi)$, depend on m and possibly other quantities, such as n_i, the sample size for the ith small area, which may increase with m. In fact, in many small-area applications, the n_i's are random outcomes, although extension to such a case is relatively straightforward (e.g. Jiang 2001); thus, for the sake of simplicity, the n_i are considered as non-random in this paper. Also, dependency of different quantities on m is suppressed for notational simplicity [e.g., $d(\psi)$ instead of $d_m(\psi)$]. The limiting process is $m \to \infty$, with other quantities considered as functions of m.

Remark 1. Another seeming advantage of the proposed MSPE estimator is that the leading term, $a(y, \hat{\psi})$, is guaranteed positive. This can be seen from the definition, that is,

$$a(y, \hat{\psi}) \;=\; \mathrm{E}\{(\hat{\theta} - \theta)^2 | y\}\Big|_{\psi=\hat{\psi}}. \tag{6}$$

See, for example, Jiang *et al.* (2018, sec. 1) for discussion on importance, and difficulty, of achieving the "double goal", that is, second-order unbiasedness and positivity at the same time, for an MSPE estimator. Typically, the leading term, (6), is $O(1)$, while the remaining term, $b(\hat{\psi}) - c(\hat{\psi})$, is $O(m^{-1})$, assuming that $\hat{\psi}$ is a root-n consistent and asymptotically normal estimator. As a result, the MSPE estimator (5), as well as the Sumca estimator (see below), are positive with probability tending to one as m increases.

A special form of the leading term, $a(y, \hat{\psi})$, deserves attention. We can write

$$a(y, \psi) \;=\; \{\hat{\theta} - \mathrm{E}(\theta|y)\}^2 + \mathrm{var}(\theta|y). \tag{7}$$

When $\hat{\theta}$ is the EBP, defined as

$$\hat{\theta} \;=\; \mathrm{E}(\theta|y)|_{\psi=\hat{\psi}}, \tag{8}$$

the first term on the right side of (7) vanishes when ψ is replaced by $\hat{\psi}$, hence

$$a(y, \hat{\psi}) \;=\; \mathrm{var}(\theta|y)|_{\psi=\hat{\psi}}. \tag{9}$$

Expression (9) is especially attractive under a general linear mixed model, defined as

$$y \;=\; X\beta + Zv + e, \tag{10}$$

where X and Z are known matrices, v is a vector of random effects, and e is a vector of errors. Suppose that v and e are independent with $v \sim N(0, G)$ and $e \sim N(0, R)$, where the covariance matrices G and R depend on a vector γ of dispersion parameters. Now suppose that θ is a linear mixed effect with the expression $\theta = a'\beta + b'v$, where a, b are

known vectors. Then, by the properties of multivariate normal distribution, $\mathrm{var}(\theta|y)$ does not depend on y; in other words, $\mathrm{var}(\theta|y)$ is a function of γ only (which is part of ψ). Write $\mathrm{var}(\theta|y) = V(\gamma)$. Then, by (9), we have

$$a(y, \hat{\psi}) \;\; = \;\; V(\hat{\gamma}). \tag{11}$$

An advantage of expression (11) is *stability*, because, typically, the variance of a smooth function of $\hat{\gamma}$ is $O(m^{-1})$. In this case, the variance of the leading term of the MSPE estimator, (5), is $O(m^{-1})$. The stability of our proposed MSPE estimator (see below) is also demonstrated numerically in our simulation studies in Section 5 as well as in the Supplementary Material. See further discussion in Section 6.

The lower-order term in (5), $b(\hat{\psi}) - c(\hat{\psi})$, corresponds to a bias correction to the leading term. Deriving an analytic form of this term, up to the order $o(m^{-1})$, may require tedious algebra. To overcome this difficulty, we approximate this term using a Monte-Carlo method. Let P_ψ denote the distribution of y with ψ being the true parameter vector. Given ψ, one can generate y under P_ψ. Let $y_{[k]}$ denote y generated under the kth Monte-Carlo sample, $k = 1, \ldots, K$. Then, we have

$$b(\psi) - c(\psi) \;\; \approx \;\; \frac{1}{K} \sum_{k=1}^{K} \left\{ a(y_{[k]}, \psi) - a(y_{[k]}, \hat{\psi}_{[k]}) \right\}, \tag{12}$$

where $\hat{\psi}_{[k]}$ denotes $\hat{\psi}$ based on $y_{[k]}$. Write the right side of (12) as $d_K(\psi)$ (note that $y_{[k]}, k = 1, \ldots, K$ also depend on ψ). Then, a Monte-Carlo assisted MSPE estimator, which we call Sumca estimator, is given by

$$\widehat{\mathrm{MSPE}}_K = a(y, \hat{\psi}) + d_K(\hat{\psi}) = a(y, \hat{\psi}) + \frac{1}{K} \sum_{k=1}^{K} \left\{ a(y_{[k]}, \hat{\psi}) - a(y_{[k]}, \hat{\psi}_{[k]}) \right\}, \tag{13}$$

where $y_{[k]}, k = 1, \ldots, K$ are generated as above with $\psi = \hat{\psi}$, and $\hat{\psi}_{[k]}$ is, again, the estimator of ψ based on $y_{[k]}$. It can be shown (see below) that the Sumca estimator is second-order unbiased with respect to the joint distribution of the data and Monte-Carlo sampling.

Remark 2. Sumca is computationally much less intensive than McJack (Jiang *et al.* 2018). This is because McJack requires $m^2/K \to 0$, while Sumca does not have such a restriction. In fact, as far as second-order unbiasedness is concerned, there is no restriction on K for Sumca; however, it is typically required that K is large enough so that the variance of the second term on the right side of (13) is of lower order than that of the first term. For example, if $\mathrm{var}\{a(y,\hat{\psi})\} = O(1)$, all that is required is $K \to \infty$. In Section 6, it is recommended that $K = m$ in standard situations. Also, for McJack, one needs to evaluate $m+1$ terms using Monte-Carlo, while for Sumca there is only one term that needs Monte-Carlo. Overall, the amount of computation for Sumca is somewhere between $1/m^3$ to $1/m^2$ of that for McJack, which could be a highly significant saving if m is relatively large.

A rigorous justification is given below, with technical details deferred to the Supplement. We first state some general conditions under which the second-order unbiasedness of the MSPE estimator (5) holds. Let $\hat{\psi}$ be the estimator of ψ, and Ψ the parameter space for ψ. Following Jiang *et al.* (2002) and Das, Jiang and Rao (2004), we consider a sequence of subsets, $\Psi_m \subset \Psi$, such that Ψ_m is compact and lies strictly in Ψ^o, the interior of Ψ, and approaches Ψ as $m \to \infty$ in the sense that $\Psi_m \subset \Psi_{m+1}$ for any $m \geq 1$, and any point in Ψ^o will be covered by Ψ_m for sufficiently large m. A truncated estimator (e.g., Das *et al.* 2004) is defined as $\hat{\psi} = \hat{\psi}_{\mathrm{o}}$ if $\hat{\psi}_{\mathrm{o}} \in \Psi_m$, and $\hat{\psi} = \psi_*$ otherwise, where ψ_* is a known vector that belongs to Ψ_m. Note that, if the true $\psi \in \Psi^o$ and $\hat{\psi}_{\mathrm{o}}$ is a consistent estimator, then we have with probability tending to one that $\hat{\psi} = \hat{\psi}_{\mathrm{o}}$; in other words, asymptotically, the truncated estimator is equal to the original estimator. In practice, the truncation does not change the value of the estimator (see a note in the first paragraph of the Supplement).

Theorem 1. The MSPE estimator given by (5) is second-order unbiased, that is, $\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{MSPE} + o(m^{-1})$ provided that the following hold:

A1. $\mathrm{E}(\theta^2|y)$ is finite almost surely, and the expectations $\mathrm{E}\{a(y,\psi)\}$ and $\mathrm{E}\{a(y,\hat{\psi})\}$ exist.

A2. $\psi \in \Psi_m$ for large m, and we have the following expansion, $d(\psi) = m^{-1}q(\psi) + r(\psi)$

for some $q(\cdot)$, $r(\cdot)$ satisfying $\sup_{\tilde{\psi}\in\Psi_m}|r(\tilde{\psi})| = o(m^{-1})$ and $\mathrm{E}\{|q(\hat{\psi}) - q(\psi)|\} = o(1)$.

Proof: We need to make the heuristic derivation below (5) for the second-order unbiasedness rigorous. Given *A1*, we just have to show that $\mathrm{E}\{d(\hat{\psi}) - d(\psi)\} = o(m^{-1})$. Note that, by *A2*, we have $d(\hat{\psi}) - d(\psi) = m^{-1}\{q(\hat{\psi}) - q(\psi)\} + r(\hat{\psi}) - r(\psi)$. Note that $\psi \in \Psi_m$ for large m by *A2*; also $\hat{\psi} \in \Psi_m$ by the definition of the truncated estimator. Thus, we have $|r(\hat{\psi}) - r(\psi)| \leq 2\sup_{\tilde{\psi}\in\Psi_m}|r(\psi)|$ for large m, and the latter is $o(m^{-1})$ by *A2*. It follows that $|\mathrm{E}\{d(\hat{\psi}) - d(\psi)\}| \leq m^{-1}\mathrm{E}(|q(\hat{\psi}) - q(\psi)|) + o(m^{-1}) = o(m^{-1})$.

Theorem A.1 in the Supplementary Material provides verifiable technical conditions under which the assumptions of Theorem 1, especially *A2*, hold. To provide some illustration on what kinds of situations where one may expect these technical conditions to hold, consider the following examples below. The first example is the Fay-Herriot model of Section 4.1 below. The mixed effect of interest is the small area mean, $\theta_i = x_i'\beta + v_i$, where $v_i \sim N(0, A)$ $(1 \leq i \leq m)$. Suppose that the Prasad-Rao estimator of A is used. Then, the technical conditions of Theorem A.1 are satisfied provided that there are positive constants b, B such that $b \leq D_i \leq B, 1 \leq i \leq m$, and $\liminf_{m\to\infty}\lambda_{\min}(m^{-1}\sum_{i=1}^m x_i x_i') > 0$, where $\lambda_{\min}$ denotes the smallest eigenvalue. The second example is a special case of the mixed logistic model of Section 4.3 in the sequel with $x_{ij}'\beta = \beta$. The mixed effect of interest is the conditional probability, $\theta_i = \mathrm{P}(y_{ij} = 1|v_i)$. Suppose that the MLE of (β, A) is used. Then, the technical conditions of Theorem A.1 are satisfied provided that $n_i, 1 \leq i \leq m$ are bounded and $\liminf_{m\to\infty}\{(\log m)^L p_m\} > 0$ for some constant $L > 0$, where p_m is the proportion of $n_i, 1 \leq i \leq m$ satisfying $n_i > 1$.

Now let us consider the Sumca estimator (13). We assume that the Monte-Carlo (MC) samples, under ψ, are generated by first generating some standard [e.g., $N(0, 1)$] random variables, say, ξ, whose distribution does not depend on ψ. We then combine ξ with ψ to produce the MC samples under ψ. A precise statement of this assumption is the following: For any given ψ, the summand in (12), $a(y_{[k]}, \psi) - a(y_{[k]}, \hat{\psi}_{[k]})$, where $\hat{\psi}_{[k]}$ is $\hat{\psi}$ based on

$y_{[k]}$, can be expressed as a function of $\xi_{[k]}$ and ψ, say, $\Delta(\xi_{[k]}, \psi)$, where $\xi_{[k]}, 1 \le k \le K$ are i.i.d. whose distribution does not depend on ψ, and are independent with the data y. This implies that the summands of (13) can be expressed as $\Delta(\xi_{[k]}, \hat{\psi}), 1 \le k \le K$ with the same $\xi_{[k]}, 1 \le k \le K$. For example, under the Fay-Herriot model (see Sections 4.1, 5.1 below), the y_i's are generated by first generating the ξ_i's and η_i's, which are independent $N(0,1)$, and then letting $y_i = x_i'\beta + \sqrt{A}\xi_i + \sqrt{D_i}\eta_i$, where $\psi = (\beta', A)'$. More generally, if the distributions of the random effects and errors are scale families, then the above assumption about the MC samples holds. As an example beyond the scale family, suppose that, in the Fay-Herriot model, the cumulative distribution function (cdf) of v_i is $F_\psi(\cdot)$ and that of e_i is $G_{i,\psi}(\cdot)$ such that the cdf's are continuous. Then, given ψ, v_i and e_i can be generated by $F_\psi^{-1}(U_i)$, $G_{i,\psi}^{-1}(V_i)$, where U_i and V_i are independent Uniform$[0,1]$ random variables, and $F_\psi^{-1}(\cdot)$, $G_{i,\psi}^{-1}(\cdot)$ are the inverse distribution functions (e.g., Jiang 2010, Theorem 7.1).

Theorem 2. Suppose that the MC samples are generated as described above, where ξ is independent with y, the original data. Then, under the assumptions of Theorem 1, we have

$$\mathrm{E}(\widehat{\mathrm{MSPE}}_K) \;=\; \mathrm{MSPE} + o(m^{-1}), \tag{14}$$

where the expectation is with respect to the joint distribution of y and ξ.

The proof of Theorem 2 is given in the Supplementary Material.

3 Post model selection predictor

In some applications, model selection is involved prior to computing the predictor. Let $\hat{\theta}_M$ be the predictor of θ computed under model M. The PMS predictor of θ is defined as $\hat{\theta}_{\mathrm{P}} = \hat{\theta}_{\hat{M}}$, where the subscript P stands for PMS, and $\hat{M}$ is the model selected by applying some model selection procedure, such as the information criteria or the fence methods (e.g., Jiang *et al.* 2008; Müller, Scealy and Welsh 2013).

We restrict our attention to consistent model selection procedures. Some restriction on the convergence rate of the model selection consistency will be imposed. Furthermore, we follow the classical setting of consistency in model selection, in which the space of candidate models, $\mathcal{M}$, is finite, and there is a unique optimal model, $M_{\rm o} \in \mathcal{M}$, that is, a true model that cannot be simplified. For example, in regression variable selection, suppose that the true (predictor) variables are a subset of the candidate variables. Then, any subset of candidate variables that include the true variables corresponds to a true model, with the understanding that the regression coefficients of the variables other than the true variables are zero. It is clear that the optimal model in this case corresponds to the subset that contains only the true variables. Mathematically, the precise definition of $M_{\rm o}$, is the limit of convergence of the model selection procedure, $\hat{M}$, that we deal with in Theorem 3 and Theorem 4 so that the conditions of those theorems are satisfied. We also assume that there is a full model, $M_{\rm f}$, in $\mathcal{M}$ so that every candidate model is a sub-model of $M_{\rm f}$. For example, in the case of regression variable selection, $M_{\rm f}$ is the subset of all of the candidate variables. It should also be made clear that these candidate models, $\mathcal{M}$, including $M_{\rm o}$ and $M_{\rm f}$, as well as the parameter spaces under these models (see below), are finite-dimensional, whose dimensions do not change as the sample size increases. Note that, under the above assumptions, $M_{\rm f}$ is, at least, a true model; however, it may not be the most efficient one in that it may allow further simplification, for example, by dropping the variables whose coefficients are zero in the regression variable selection problem. Finding the most efficient true model, that is, $M_{\rm o}$, is often the purpose of model selection.

For $M \in \mathcal{M}$, let ψ_M denote the vector of parameters under M, $\hat{\psi}_M$ the estimator of ψ_M, and $\hat{\theta}_M$ the predictor of θ, the mixed effect of interest, derived under M. If M is given, then $\hat{\psi}_M$ and $\hat{\theta}_M$ are the ones discussed in the previous sections. The only additional step, here, is the process of model selection prior to computing $\hat{\psi}_M$ and $\hat{\theta}_M$. Let $\hat{M}$ denote the selected model via the model selection procedure. Then, the PMS estimator of the

parameters, under the selected model, is $\hat{\psi}_{\rm P} = \hat{\psi}_{\hat{M}}$; the PMS predictor of θ is $\hat{\theta}_{\rm P} = \hat{\theta}_{\hat{M}}$. A measure of uncertainty for $\hat{\theta}_{\rm P}$ is, again, its MSPE defined as $\mathrm{MSPE}(\hat{\theta}_{\rm P}) = \mathrm{E}(\hat{\theta}_{\rm P} - \theta)^2$. Note that the MSPE is a measure of overall uncertainty in $\hat{\theta}_{\rm P}$, which includes the uncertain in model selection, in parameter estimation given the selected model, and prediction under the selected model. It is not necessarily true that the overall uncertainty is larger than uncertainty in prediction under a fixed model. For example, if a simpler model is selected, it may help to reduce the uncertainty in parameter estimation; on the other hand, if a more complex model is selected, it may help to reduce the prediction error. The bottom line is that the MSPE of $\hat{\theta}_{\rm P}$ is more complex than that of $\hat{\theta}_M$ under a fixed M.

The Sumca method can be used to estimate the MSPE of $\hat{\theta}_{\rm P}$. First note that, in (2), E stands for the true conditional expectation. Following Jiang *et al.* (2015), in a PMS situation, the conditional expectations in (2) is evaluated under $M_{\rm f}$, which is a true model, and $\psi_{\rm f} \equiv \psi_{M_{\rm f}}$, the vector of true parameters under $M_{\rm f}$, if the latter is given. With this understanding, all of the derivations in Section 2 carry through with ψ replaced by $\psi_{\rm f}$, and $\hat{\psi}$ replaced by $\hat{\psi}_{\rm f} \equiv \hat{\psi}_{M_{\rm f}}$, the estimator of $\psi_{\rm f}$. Note that there is no need to change the notation E, because expectation under $M_{\rm f}$ (and $\psi_{\rm f}$) is the true expectation.

Two facts need to be noted. First, let $\psi_{\rm o} \equiv \psi_{M_{\rm o}}$ denote the true parameter vector under $M_{\rm o}$. Then, because $M_{\rm f}$ is also a true model, any expectation under $M_{\rm o}$ and $\psi_{\rm o}$ is the same as that under $M_{\rm f}$ and $\psi_{\rm f}$. As a result, any function of $\psi_{\rm o}$ that is obtained via taking expectation (even repeatedly) can be expressed as a function of $\psi_{\rm f}$ by the same operation. As far as this paper is concerned, only such functions are involved.

Second, let $\mathrm{E}_{\rm f}$ denote expectation, or conditional expectation, under $M_{\rm f}$ and the true $\psi_{\rm f}$. We assume that the mixed effect of interest, θ, is the same under any true model and true parameters. For example, if $\theta = x_i'\beta + v_i$, the value of θ does not change if one adds, or drops, zero components to β (and adjusts x_i correspondingly). Then, the PMS version of

(2) can be written as

$$a(y,\psi_{\rm f}) \;=\; \hat{\theta}_{\rm P}^2 - 2\hat{\theta}_{\rm P}h_1(y,\psi_{\rm f}) + h_2(y,\psi_{\rm f}), \tag{15}$$

where $h_j(y,\psi_{\rm f}) = {\rm E}_{\rm f}(\theta^j|y)$, $j=1,2$. Let $a_{\rm o}(y,\psi_{\rm f})$, $a_M(y,\psi_{\rm f})$ be (15) with $\hat{\theta}_{\rm P}$ replaced by $\hat{\theta}_{\rm o} \equiv \hat{\theta}_{M_{\rm o}}$, $\hat{\theta}_M$, respectively. We have with the following obvious equations:

$$a(y,\psi_{\rm f}) - a_{\rm o}(y,\psi_{\rm f}) \;=\; \sum_{M\neq M_{\rm o}} \{a_M(y,\psi_{\rm f}) - a_{\rm o}(y,\psi_{\rm f})\}1_{(\hat{M}=M)}, \tag{16}$$

$$a_M(y,\psi_{\rm f}) - a_{\rm o}(y,\psi_{\rm f}) \;=\; \hat{\theta}_M^2 - \hat{\theta}_{\rm o}^2 - 2h_1(y,\psi_{\rm f})(\hat{\theta}_M - \hat{\theta}_{\rm o}), \tag{17}$$

using (15) for (17). Let $b_{\rm o}(\psi_{\rm f}) = {\rm E}_{\rm f}\{a_{\rm o}(y,\psi_{\rm f})\}$ and $c_{\rm o}(\psi_{\rm f}) = {\rm E}_{\rm f}\{a_{\rm o}(y,\hat{\psi}_{\rm f})\}$. An alternative expression of the MSPE estimator, (5) (note that now $\hat{\psi} = \hat{\psi}_{\rm f}$), that is useful for the theoretical derivation (but not for computing purposes) is the following: Define $\widehat{\rm MSPE}_1 = a_{\rm o}(y,\hat{\psi}_{\rm f}) + b_{\rm o}(\hat{\psi}_{\rm f}) - c_{\rm o}(\hat{\psi}_{\rm f})$; then, we have

$$\widehat{\rm MSPE} - \widehat{\rm MSPE}_1 \;=\; 2{\rm E}_{\rm f}[\{h_1(y,\hat{\psi}_{\rm f}) - h_1(y,\psi_{\rm f})\}(\hat{\theta}_{\rm P} - \hat{\theta}_{\rm o})1_{(\hat{M}\neq M_{\rm o})}]\Big|_{\psi_{\rm f}=\hat{\psi}_{\rm f}}. \tag{18}$$

Note that $\widehat{\rm MSPE}_1$ looks quite similar to the $\widehat{\rm MSPE}$ in (5), with one difference: Typically, $\hat{\theta}$ involves a parameter estimator. In (5), the parameter estimator in $\hat{\theta}$ is the same as that involved elsewhere, that is, $\hat{\psi}$. In $\widehat{\rm MSPE}_1$, the parameter estimator involved in $\hat{\theta}_{\rm o}$ is $\hat{\psi}_{\rm o}$; elsewhere it is $\hat{\psi}_{\rm f}$. The most important thing is that there is no model selection involved in $\widehat{\rm MSPE}_1$ so that its second-order unbiasedness can be proved similarly. The remaining term, that is, the right side of (18), is negligible as long as $\hat{M}$ has good consistency property, which is made precise by condition (iii) in Theorem 3 below.

We can now apply Theorem 1 to the current PMS case. Note that $\widehat{\rm MSPE}$ is now (5) with $\hat{\psi}$ replaced by $\hat{\psi}_{\rm f}$. For a random variable ξ, define $\|\xi\|_2 = \{{\rm E}(\xi^2)\}^{1/2}$. Following Jiang *et al.* (2002) (also Das *et al.* 2004), to ensure the existence of the MSPE we consider a truncated version of $\hat{\psi}_{\rm f}$. Let $\Psi_{{\rm f},m}$ be a compact subspace of $\Psi_{\rm f}$, the parameter space of $\psi_{\rm f}$, which is expanding as m increases (for example, $\Psi_{{\rm f},m} = \{\psi_{\rm f} : |\psi_{\rm f}| \leq b_m\}$, where $|\cdot|$

denotes the Euclidean norm and $b_m \to \infty$ as $m \to \infty$). Note that such a truncation is only used in a (rigorous) theoretical argument; there is no need to truncate an estimator in practice. For example, one can change the b_m in the definition of $\Psi_{\mathrm{f},m}$ by cb_m, where c is any positive constant, and the same argument goes through; on the other hand, it may be argued that the $\hat{\psi}_{\mathrm{f}}$ obtained from the real data satisfies $\hat{\psi}_{\mathrm{f}} \in \Psi_{\mathrm{f},m}$ for some constant c.

Theorem 3. Suppose that the conditions of Theorem 1 hold with ψ replaced by ψ_{f}. Furthermore, suppose that the following hold: (i) $\sup_{\psi_{\mathrm{f}} \in \Psi_{\mathrm{f},m}} \|h_1(y, \hat{\psi}_{\mathrm{f}}) - h_1(y, \psi_{\mathrm{f}})\|_2 \leq cm^{\delta}$ for some constant $c, \delta > 0$; (ii) $\|\hat{\theta}_{\mathrm{o}} - \theta\|_2 = O(m^{\delta})$; and (iii) $\|\hat{\theta}_{\mathrm{P}} - \hat{\theta}_{\mathrm{o}}\|_2 = o(m^{-1-\delta})$. Then, we have $\mathrm{E}(\widehat{\mathrm{MSPE}}) = \mathrm{MSPE} + o(m^{-1})$. Define $\hat{\psi}_{\mathrm{f}}$ as ψ_{f}^*, where ψ_{f}^* is a known vector in $\Psi_{\mathrm{f},m}$, if $\hat{\psi}_{\mathrm{f}} \notin \Psi_{\mathrm{f},m}$.

Note that condition (iii) is related to the convergence rate of $\mathrm{P}(\hat{M} = M_{\mathrm{o}})$ going to one, or $\mathrm{P}(\hat{M} \neq M_{\mathrm{o}})$ going to zero. To see this, suppose that $\hat{\theta}_{\mathrm{o}}, \hat{\theta}_{\mathrm{P}}$ satisfy $\|\hat{\theta}_{\mathrm{P}} - \hat{\theta}_{\mathrm{o}}\|_4 = O(m^{\delta})$, where $\|\xi\|_4 = \{\mathrm{E}(\xi^4)\}^{1/4}$, then by the Cauchy-Schwarz inequality we have

$$\mathrm{E}(\hat{\theta}_{\mathrm{P}} - \hat{\theta}_{\mathrm{o}})^2 = \mathrm{E}\{(\hat{\theta}_{\mathrm{P}} - \hat{\theta}_{\mathrm{o}})^2 1_{(\hat{M} \neq M_{\mathrm{o}})}\} \leq O(m^{2\delta})\mathrm{P}(\hat{M} \neq M_{\mathrm{o}})^{1/2}.$$

Thus, provided that $\mathrm{P}(\hat{M} \neq M_{\mathrm{o}}) = o(m^{-4-8\delta})$, the above inequality implies that $\mathrm{E}(\hat{\theta}_{\mathrm{P}} - \hat{\theta}_{\mathrm{o}})^2 = o(m^{-2-2\delta})$, hence $\|\hat{\theta}_{\mathrm{P}} - \hat{\theta}_{\mathrm{o}}\|_2 = o(m^{-1-\delta})$. It is know that, in model selection, the convergence rate of $\mathrm{P}(\hat{M} \neq M_{\mathrm{o}}) \to 0$ can be at the exponential rate, which is much faster than $m^{-\lambda}$ for any $\lambda > 0$ (see Section A.1.4 of the Supplementary Material for an example).

The proof of Theorem 3 is given in the Supplementary Material.

Furthermore, Theorem A.1 in the Supplementary Material can be extended to the PMS case (see Theorem 5 of Jiang and Torabi 2019).

The next result is a PMS version of Theorem 2. Here, the Sumca estimator, $\widehat{\mathrm{MSPE}}_K$, is given by (13) with ψ replaced by ψ_{f} (and $\hat{\psi}$ by $\hat{\psi}_{\mathrm{f}}$).

Theorem 4. Suppose that the MC samples are generated as described above Theorem 2, where ξ is independent with y. Then (14) holds under the assumption of Theorem 3,

where the expectation in (14) is with respect to the joint distribution of y and ξ.

The proof of Theorem 4 is similar to that of Theorem 2, and therefore omitted.

4 Examples

In this section, we demonstrate the Sumca estimator with a few examples. The examples include a popular area-level SAE model, a situation of PMS SAE, and a mixed logistic model used for estimating small area proportions.

4.1 Fay-Herriot model

A widely used SAE model is the Fay-Herriot model (Fay and Herriot 1979). The model can be expressed in terms of a linear mixed model:

$$y_i \;\; = \;\; x_i'\beta + v_i + e_i, \;\; i = 1, \ldots, m, \tag{19}$$

where y_i is a direct estimator from the ith area, x_i is a vector of known covariate, β is a vector of unknown fixed effects, v_i is an area-specific random effect, and e_i is a sampling error. It is assumed that $v_i, e_i, 1 \leq i \leq m$ are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$, where A is an unknown variance but D_i is assumed known, $1 \leq i \leq m$. Here, the vector of unknown parameters involved in the model is $\psi = (\beta', A)'$. As noted earlier (see the paragraph following the proof of Theorem 1), the technical conditions of Theorem A.1 in the Supplement, under the Fay-Herriot model, are satisfied provided that the D_i's are bounded and bounded away from zero, and $\liminf_{m\to\infty} \lambda_{\min}(m^{-1}\sum_{i=1}^m x_i x_i') > 0$.

Suppose that we are interested in estimating the small area mean, $\theta_i = x_i'\beta + v_i$, for each small area i, using the EBP. This is a special case of (10), thus, by (9), we have

$$a_i(y, \hat{\psi}) = \mathrm{var}(\theta_i|y)|_{\psi=\hat{\psi}} = \frac{\hat{A}D_i}{\hat{A} + D_i}. \tag{20}$$

The Sumca estimator, (13), is thus given by

$$\widehat{\mathrm{MSPE}}_{i,K} = \frac{\hat{A}D_i}{\hat{A}+D_i} + \frac{1}{K}\sum_{k=1}^{K}\left\{a_i(y_{[k]},\hat{\psi}) - a_i(y_{[k]},\hat{\psi}_{[k]})\right\}. \tag{21}$$

To be more specific, let us consider the Prasad-Rao estimator of A (Prasad and Rao 1990), given by $\hat{A} = (m-p)^{-1}\{y'P_{X^\perp}y - \mathrm{tr}(P_{X^\perp}D)\}$, where $p = \mathrm{rank}(X)$, $y = (y_i)_{1\le i\le m}$, $P_{X^\perp} = I_m - P_X$, $P_X = X(X'X)^{-1}X'$ for $X = (x_i')_{1\le i\le m}$, and $D = \mathrm{diag}(D_1,\ldots,D_m)$. Given $\hat{A}$, the estimator of β is given by $\hat{\beta} = \tilde{\beta}(\hat{A})$, where $\tilde{\beta}(A) = (X'V^{-1}X)^{-1}X'V^{-1}y$ with $V = AI_m + D$. For this particular EBP, Prasad and Rao (1990) showed that the following Prasad-Rao MSPE estimator for $\hat{\theta}_i$ is second-order unbiased:

$$\begin{aligned}\widehat{\mathrm{MSPE}}_{i,\mathrm{PR}} &= \frac{\hat{A}D_i}{\hat{A}+D_i} + \left(\frac{D_i}{\hat{A}+D_i}\right)^2 x_i'\left(\sum_{j=1}^{m}\frac{x_jx_j'}{\hat{A}+D_j}\right)^{-1}x_i \\ &\quad + \frac{4D_i^2}{(\hat{A}+D_i)^3m^2}\sum_{j=1}^{m}(\hat{A}+D_j)^2.\end{aligned} \tag{22}$$

Comparing (21) with (22), it is clear that the two estimators have the same leading $O(1)$ term, $\hat{A}D_i/(\hat{A}+D_i)$; the only difference is the remaining lower-order term, which corresponds to a bias correction. Prasad and Rao (1990) used an analytic bias correction while Sumca uses a Monte-Carlo bias correction. It can be shown that, in this case, we have

$$\begin{aligned}a_i(y_{[k]},\hat{\psi}) - a_i(y_{[k]},\hat{\psi}_{[k]}) &= D_i\left(\frac{\hat{A}}{\hat{A}+D_i} - \frac{\hat{A}_{[k]}}{\hat{A}_{[k]}+D_i}\right) \\ &\quad + \left\{\left(\frac{\hat{A}_{[k]}}{\hat{A}_{[k]}+D_i} - \frac{\hat{A}}{\hat{A}+D_i}\right)y_{[k],i} + D_ix_i'\left(\frac{\hat{\beta}_{[k]}}{\hat{A}_{[k]}+D_i} - \frac{\hat{\beta}}{\hat{A}+D_i}\right)\right\}^2,\end{aligned} \tag{23}$$

where $y_{[k],i}$ is the ith component of $y_{[k]}$, and $\hat{\psi}_{[k]} = (\hat{\beta}_{[k]}', \hat{A}_{[k]})'$.

Another existing approach to second-order unbiased MSPE estimation is double (parametric) bootstrap (DB; Hall and Maiti 2006b). Because it is computationally intensive, in the simulation study, later in Section 5.1, we compare the DB approach with PR and Sumca in the case of relatively small m. To obtain the DB estimator, we generate the bootstrap

data under (19) with $\hat{\beta}, \hat{A}$ as the true parameters. The small area means for the bootstrap data are defined accordingly. We then obtain the parameter estimates and small area mean predictors based on the bootstrapped data. This leads to the first-phase (or single) bootstrapped MSPE, with the first-phase bootstrap sample size B_1. It is known (e.g., Hall and Maiti 2006a) that the first-phase bootstrapped MSPE is only first-order unbiased. Thus, in the next step, we draw a second-phase bootstrap sample from a given bootstrap sample using the bootstrapped model parameters given above. Proceeding as above to obtain the second-phase bootstrapped MSPE, with the second-phase bootstrap sample size B_2. Due to the computational cost, it is suggested (Hall and Maiti 2006a, b) that B_1 is moderately large and B_2 is smaller than B_1. Hall and Maiti (2006b) further suggests two nonnegative modifications of the second-phase bootstrapped MSPE. These are the final DB MSPE estimators, denoted by $\widehat{\mathrm{MSPE}}_{i,\mathrm{Boot1}}$ and $\widehat{\mathrm{MSPE}}_{i,\mathrm{Boot2}}$, respectively.

4.2 Area-level model with model selection

Datta, Hall and Mandal (2011; hereafter, DHM) proposed a method of model selection by testing for the presence of the area-specific random effects, v_i, in the Fay-Herriot model (19). Once again, consider the Prasad-Rao estimator of A given below (21). Then, DHM is equivalent to testing the null hypothesis $\mathrm{H}_0 : A = 0$. The test statistic, $T = \sum_{i=1}^{m} D_i^{-1}(y_i - x_i'\tilde{\beta})^2$, where $\tilde{\beta} = \tilde{\beta}(0)$ [defined above (22)], has a χ^2_{m-p} distribution under H_0. If H_0 is rejected, the EBLUP is used to estimate the small area mean θ_i; if H_0 is accepted, the estimator $\hat{\theta}_i = x_i'\tilde{\beta}$ is used to estimate θ_i, where $\tilde{\beta} = (X'D^{-1}X)^{-1}X'D^{-1}y$. Note that, the combined predictor, $\hat{\theta}_i$, is different from the traditional EBLUP; namely, we have

$$\hat{\theta}_i \;=\; \begin{cases} \hat{A}(\hat{A}+D_i)^{-1}y_i + D_i(\hat{A}+D_i)^{-1}x_i'\hat{\beta}, & \text{if } T > \chi^2_{m-p}(1-\alpha), \\ x_i'\tilde{\beta}, & \text{if } T \le \chi^2_{m-p}(1-\alpha), \end{cases} \tag{24}$$

where α is the level of significance. As for the MSPE estimator corresponding to (24), DHM proposed the following:

$$\widehat{\mathrm{MSPE}}_{i,\mathrm{DHM}} = (22) \text{ if } T > \chi^2_{m-p}(1-\alpha), \text{ and } x_i'(X'D^{-1}X)^{-1}x_i \text{ otherwise.} \tag{25}$$

On the other hand, our Sumca approach applies to any kind of predictor, thus, in particular, to the predictor (24). The corresponding Sumca estimator is given by (13), where $\hat{\beta}, \hat{A}$ are the same as in (24) (note that these are consistent estimators of β, A, respectively, regardless of the null hypothesis). Note that (9) no longer holds in this case, because $\hat{\theta}_i$ is not the EBP. In fact, by the right side of (7), we have

$$a_i(y, \hat{\psi}) = \frac{\hat{A}D_i}{\hat{A}+D_i} + \left(\hat{\theta}_i - \frac{\hat{A}}{\hat{A}+D_i}y_i - \frac{D_i}{\hat{A}+D_i}x_i'\beta\right)^2, \tag{26}$$

$$\begin{aligned} & a_i(y_{[k]}, \hat{\psi}) - a_i(y_{[k]}, \hat{\psi}_{[k]}) \\ = \; & D_i\left(\frac{\hat{A}}{\hat{A}+D_i} - \frac{\hat{A}_{[k]}}{\hat{A}_{[k]}+D_i}\right) + \left(\hat{\theta}_i - \frac{\hat{A}}{\hat{A}+D_i}y_{[k],i} - \frac{D_i}{\hat{A}+D_i}x_i'\hat{\beta}\right)^2 \\ & - \left(\hat{\theta}_i - \frac{\hat{A}_{[k]}}{\hat{A}_{[k]}+D_i}y_{[k],i} - \frac{D_i}{\hat{A}_{[k]}+D_i}x_i'\hat{\beta}_{[k]}\right)^2. \end{aligned} \tag{27}$$

4.3 A mixed logistic model

Jiang and Lahiri (2001) considered the following mixed logistic model for SAE with binary data. Suppose that, given the area-specific random effects, $v_1, \ldots, v_m$, binary responses $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i$ are conditionally independent such that $\mathrm{P}(y_{ij} = 1|v) = p_{ij}$ with $\mathrm{logit}(p_{ij}) = x_{ij}'\beta + v_i$, where $v = (v_i)_{1\le i\le m}$, x_{ij} is a vector of known covariates, and β is a vector of unknown fixed effects. Furthermore, it is assumed that $v_1, \ldots, v_m$ are independent and distributed as $N(0, A)$, where A is an unknown variance. For simplicity, consider the case that x_{ij} is at area level, that is, $x_{ij} = x_i$. The mixed effect of interest is the conditional probability, $\theta_i = h(x_i'\beta + v_i)$, where $h(u) = e^u/(1+e^u)$. As noted earlier (see the paragraph following the proof of Theorem 1), for the special case of

the mixed logistic model with $x_i'\beta = \beta$, the technical conditions of Theorem A.1 in the Supplement are satisfied provided that the n_i's are bounded and $\liminf_{m\to\infty}\{(\log m)^L p_m\} > 0$ for some $L > 0$, where $p_m = m^{-1}\#\{1 \le i \le m : n_i > 1\}$.

It is easy to show that $f(v|y) = \prod_{i=1}^m f(v_i|y_i)$, where $y_i = (y_{ij})_{1\le j\le n_i}$ and, with a little abuse of the notation, $f(\xi|\eta)$ denotes the conditional pdf of ξ given η. It follows that

$$h_{i,s}(y,\psi) = \mathrm{E}(\theta_i^s|y) = \frac{\int\{h(x_i'\beta + v_i)\}^s f_\beta(y_i|v_i) f_A(v_i) dv_i}{\int f_\beta(y_i|v_i) f_A(v_i) dv_i}, \quad s = 1, 2, \tag{28}$$

where $f_\beta(y_i|v_i) = \exp\left\{y_{i\cdot}(x_i'\beta + v_i) - n_i \log(1 + e^{x_i'\beta + v_i})\right\}$ denote the conditional pmf of y_i given v_i, with $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$, $\psi = (\beta', A)'$, and $f_A(\cdot)$ denotes the pdf of $N(0, A)$. Expressions (28) involve some one-dimensional integrals, which can be evaluated numerically [e.g., using the R function integrate()]. Given the expressions (28), the EBP of θ_i is given by $\hat\theta_i = h_{i,1}(y, \hat\psi)$, where $\hat\psi$ is the maximum likelihood estimator (MLE) of ψ, which can be computed numerically by fitting the mixed logistic model which is a special case of the generalized linear mixed models (GLMMs; e.g., Jiang 2007).

Our Sumca estimator also applies to this case. Note that, because $\hat\theta_i$ is the EBP based on the same $\hat\psi$, the simplified expression (9) holds, leading to $a_i(y, \hat\psi) = h_{i,2}(y, \hat\psi) - \hat\theta_i^2$, $a_i(y_{[k]}, \hat\psi_{[k]}) = h_{i,2}(y_{[k]}, \hat\psi_{[k]}) - \hat\theta_{i,[k]}^2$, where $\hat\theta_{i,[k]} = h_{i,1}(y_{[k]}, \hat\psi_{[k]})$; however, there is no simplification for $a_i(y_{[k]}, \hat\psi)$.

As a comparison, Jiang *et al.* (2002; JLW) proposed a jackknife method for estimating the MSPE, which can be expressed as

$$\widehat{\mathrm{MSPE}}_{i,\mathrm{JLW}} = B_i(\hat\psi) - \frac{m-1}{m}\sum_{i'=1}^m \{B_i(\hat\psi_{-i'}) - B_i(\hat\psi)\} + \frac{m-1}{m}\sum_{i'=1}^m (\hat\theta_{i,-i'} - \hat\theta_i)^2. \tag{29}$$

In (29), $B_i(\psi)$ is the MSPE of the BP, $\tilde\theta_i = h_{i,1}(y, \psi)$, where ψ is the true parameter vector. Note that $h_{i,s}(y, \psi)$ depends on y only through $y_{i\cdot}$, therefore can be written as $h_{i,s}(y_{i\cdot}, \psi)$, $s = 1, 2$. Furthermore, given v_i, $y_{i\cdot}$ is distributed as $\mathrm{Binomial}(n_i, \theta_i)$. It can be shown that

$$B_i(\psi) = \sum_{k=1}^{n_i} C_k^{n_i}\left[h_{i,2}(k,\psi) - \{h_{i,1}(k,\psi)\}^2\right]\mathrm{E}\{\theta_i^k(1-\theta_i)^{n_i-k}\}, \tag{30}$$

where C_k^n is the binomial coefficient of n choose k, $\theta_i = h(x_i'\beta + v_i)$, and the expectation is taken with respect to $v_i \sim N(0, A)$. Again, $B_i(\psi)$ can be evaluated via numerical integration. The $\hat{\psi}_{-i'}$ in (29) is obtained the same way as $\hat{\psi}$ but with the i'th cluster of the data, $(y_{i'}, x_{i'})$, deleted; and $\hat{\theta}_{i,-i'} = h_{i,1}(y_{i\cdot}, \hat{\psi}_{-i'})$.

Another alternative approach to Sumca is parametric bootstrap. Since the DB approach (see Section 4.1) is computationally too intensive in this case, single bootstrap was considered in our simulation study, reported in a technical report (Jiang and Torabi 2019).

5 Simulation studies

We carry out simulation studies to compare performance of the Sumca MSPE estimator with popular existing methods. More simulation studies are deferred to Section A.2 of the Supplementary Material as well as the online technical report (Jiang and Torabi 2019).

5.1 Fay-Herriot model

We consider the following Fay-Herriot model: $y_i = \beta_0 + \beta_1 x_i + v_i + e_i$, $i = 1, ..., m$, where $m = 2m_1$, $D_i = D_{i1}$ for $1 \le i \le m_1$ and $D_i = D_{i2}$ for $m_1 + 1 \le i \le m$. We choose $\beta_0 = \beta_1 = 1, A = 10$. The D_{i1} are generated from the Uniform$[3.5, 4.5]$ distribution; D_{i2} are generated from the Uniform$[0.5, 1.5]$ distribution. The x_i's are generated from the Uniform$[0, 1]$ distribution. The x_i's and D_i's are fixed during the simulation study.

We consider three different sample sizes $m = 20, 50, 200$. We run $R = 1,000$ simulations to calculate $\widehat{\mathrm{MSPE}}_{i,K}$, $\widehat{\mathrm{MSPE}}_{i,\mathrm{PR}}$, $\widehat{\mathrm{MSPE}}_{i,\mathrm{Boot1}}$ and $\widehat{\mathrm{MSPE}}_{i,\mathrm{Boot2}}$. See Section 4.1 for the method descriptions. For the Sumca estimator, we use $K = m \vee 100$. As for the bootstrap methods, we consider $B_1 = 1,000$ and $B_2 = 100$.

To evaluate performance of the MSPE estimators, we calculate the empirical MSPE (EMSPE) of $\hat{\theta}_i^{(r)}$ over the simulation runs, where $\hat{\theta}_i^{(r)}$ is the EBP of the true small area

mean, $\theta_i^{(r)}$, for the rth simulation run, $r = 1, \ldots, R$. Specifically, we have

$$\text{EMSPE} \quad = \quad \frac{1}{R}\sum_{r=1}^{R}\{\hat{\theta}_i^{(r)} - \theta_i^{(r)}\}^2, \qquad i = 1, ..., m; \tag{31}$$

The percentage relative bias (% RB) of a MSPE estimator, $\widehat{\text{MSPE}}$, is defined as

$$\%\ \text{RB} \quad = \quad 100 \times [\{\text{E}(\widehat{\text{MSPE}}) - \text{EMSPE}\}/\text{EMSPE}], \tag{32}$$

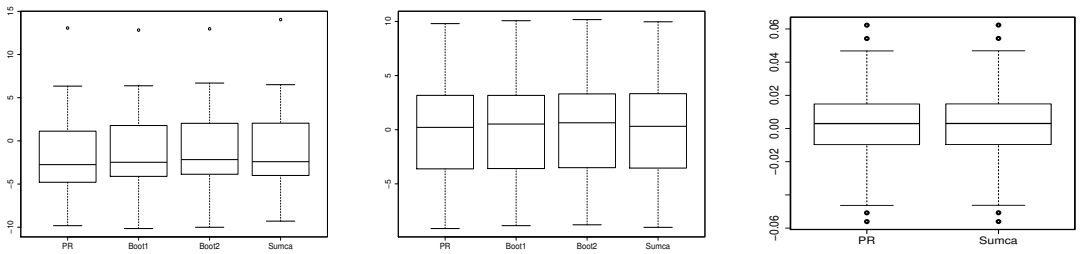
where $\text{E}(\widehat{\text{MSPE}})$ is the average of the simulated $\widehat{\text{MSPE}}$.

Figure 1 shows that all four methods perform very well in terms of %RB for different number of small areas. Comparatively, the bias of the Prasad-Rao estimator tend to be more negative, indicating (slight) underestimation, while that of the Sumca estimator tend to be more positive, indicating (slight) overestimation. See Figure A.2 of the Supplementary Material for illustration. Especially for $m = 50$, the median %RB of PR and Sumca methods seem to be closer to zero than the DB methods. Note that we only compare with DB for $m = 20$ and 50, as DB gets computationally more intensive for larger m.

Following Slud and Maiti (2006), we also consider variability of the MSPE estimators for different m. It was found that Sumca estimator is very stable for different m. See Section A.2 of the Supplementary Material for detail.

It should be noted that the Prasad-Rao MSPE estimator is specifically derived to incorporate the Prasad-Rao estimator of A (Prasad and Rao 1990); on the other hand, our unified Sumca approach applies to any procedure of parameter estimation, including the special case of Prasad-Rao estimator. Furthermore, our simulation results show that, when it comes to this special case, Sumca is as competitive as the method specifically designed for the case. Again, see Section A.2 of the Supplementary Material for more detail.

Finally, although we have considered relatively small number of replications ($K = m \vee 100$) for the Sumca estimator (see the next section for discussion), the values of the Sumca estimators were positive in all of the simulation runs, another desirable property for an MSPE estimator (see Remark 1 in Section 2).



(a) (b) (c)

Figure 1: Boxplots of % RB of MSPE estimates using Sumca, DB (Boot1 and Boot2) and PR methods: (a) $m = 20$; (b) $m = 50$; (c) $m = 200$.

5.2 Area-level model with model selection

We carry out a simulation study to compare performance of the Sumca estimator with the DHM estimator in a PMS situation (see Section 3). As described earlier, DHM proposed model selection by testing for the presence of the area-specific random effects or, equivalently, $\mathrm{H}_0 : A = 0$ before determining which SAE method to use (see Section 4.2).

The data are generated under the following model: $y_i = \beta_{01}+\beta_1 x_i+v_i+e_i$, $1 \le i \le m_1$; $y_i = \beta_{02} + \beta_1 x_i + v_i + e_i$, $m_1 + 1 \le i \le m$, where the v_i's and e_i's satisfy the assumptions of the Fay-Herriot model, $D_i = D_{i1}$ for $1 \le i \le m_1$ and $D_i = D_{i2}$ for $m_1 + 1 \le i \le m$ with $m = 2m_1$. The parameters β_{01}, β_{02} are to be determined later, $\beta_1 = 1$, and three different true values of A are considered: $A = 0, 0.5, 1$. The D_{i1} are generated from the Uniform$[0.5, 1.5]$ distribution and D_{i2} are from the Uniform$[15.5, 16.5]$ distribution. The x_i's are generated from the Uniform $[0, 1]$ distribution. The D_i's and x_i's are fixed throughout the simulation study.

The model we are fitting, however, is the standard Fay-Herriot model (19) with $x_i'\beta = \beta_0 + \beta_1 x_i$, where the x_i, D_i are as above, and β_0, β_1, A are unknown fixed effects. In other words, there is a potential model misspecification (in assuming that $\beta_{01} = \beta_{02} = \beta_0$),

but we pretend that this is not known. Following a suggestion by Molina, Rao and Datta (2015), we consider $\alpha = 0.2$ as the level of significance in testing $A = 0$. Other than the hypothesis testing, the SAE method is the same as in the previous subsection; in particular, we consider the Prasad-Rao estimators of A and β, as described below (21).

We first consider a case where there is no model misspecification, that is, $\beta_{01} = \beta_{02} = \beta_0 = 1$. The main purpose is study performance of Sumca (and DHM) as the sample size, m, increases, when the underlying model is correct. We consider $m = 20, 50, 100$. We compare performance of Sumca and DHM methods in terms of %RB. We use $K = 100$ for the Sumca estimator. The results, based on $R = 1,000$ simulation runs, are illustrated by Figure 2; some summary statistics are reported in Table 1. Both the figure and table show the improved performance of Sumca as m increases, both in terms of the mean (median) and especially in term of the standard deviation of the %RB. This is what the asymptotic theory has predicted (see Theorem 4). Overall, the performance of DHM improves, too, as m increases, although, comparatively, Sumca performs much better than DHM, especially under the alternatives. The average %RB over the small areas appears to be positive for DHM, and negative for Sumca. The empirical probability of rejection in the hypothesis test is also reported in Table 1. The test appears to perform very well.

Next, we consider the case that the underlying model is misspecified, that is, $\beta_{01} = 1, \beta_{02} = 4$. It should be noted that the theory (both Theorem 3 and Theorem 4) is established under the assumption that there is no model misspecification other than the part that is subject to model selection. Thus, the asymptotic theory does not predict the behavior of Sumca in this case. Nevertheless, we can still compare the relative performance of Sumca and DHM in this case. Figure 3, which is also based on $R = 1,000$ simulation runs, shows that, in all scenarios, Sumca still performs better than DHM, especially under the alternative, although the difference is somewhat less striking compared to the case of no model misspecification. It is also seen that the difference between the two methods is more signif-

Table 1: Summary statistics for Figure 2: Probability of rejection (POR; nominal level $= 0.2$); and empirical mean (standard deviation) of %RB over different small areas

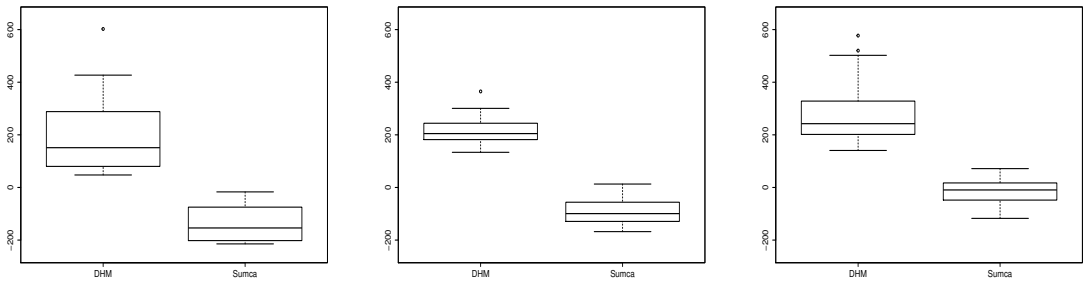
	$A = 0$			$A = 0.5$			$A = 1$		
	$m = 20$	$m = 50$	$m = 100$	$m = 20$	$m = 50$	$m = 100$	$m = 20$	$m = 50$	$m = 100$
POR	0.178	0.197	0.192	0.457	0.629	0.841	0.688	0.897	0.988
DHM	201.04	211.86	271.43	440.42	267.57	234.23	545.58	274.01	153.34
	(150.70)	(48.99)	(93.04)	(424.62)	(212.36)	(175.40)	(556.89)	(252.70)	(135.18)
Sumca	-137.77	-90.61	-16.28	-111.08	-75.99	-50.46	-98.48	-66.08	-42.75
	(67.86)	(47.09)	(44.58)	(42.39)	(9.96)	(9.17)	(31.18)	(6.43)	(7.52)

icant for smaller m than for larger m. An explanation is that both methods perform better in larger sample, although this is not implied by the asymptotic theory. The corresponding table of summary statistics (similar to Table 1) is omitted.

6 Concluding remarks and discussion

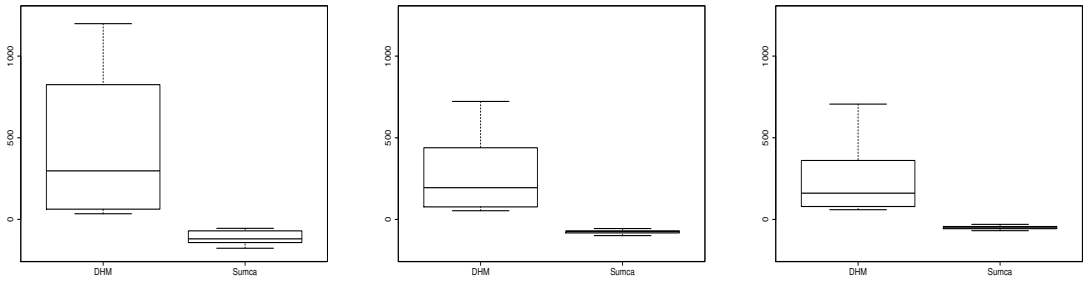
As we have demonstrated, Sumca is a general method for deriving a second-order unbiased MSPE estimator. In fact, there is a simple message that is delivered, which is even more broadly applicable: Roughly speaking, if one has a first-order unbiased MSPE estimator, and one can simulate its bias given any values of the true parameters, simulate the bias at the estimated parameters. Then, the first-order unbiased MSPE estimator, minus the bias simulated at the estimated parameters, is a second-order unbiased MSPE estimator. Importantly, the simulation of the bias does not involve double bootstrap, provided that the first-order unbiased MSPE estimator can be computed without using bootstrapping.

There is a significant computational advantage of Sumca estimator compared to existing resampling methods in SAE for obtaining second-order unbiased MSPE estimators of small area predictors. For example Hall and Maiti (2006ab) proposed the DB method to correct a bootstrap MSPE estimator in order to achieve the second-order unbiasedness.



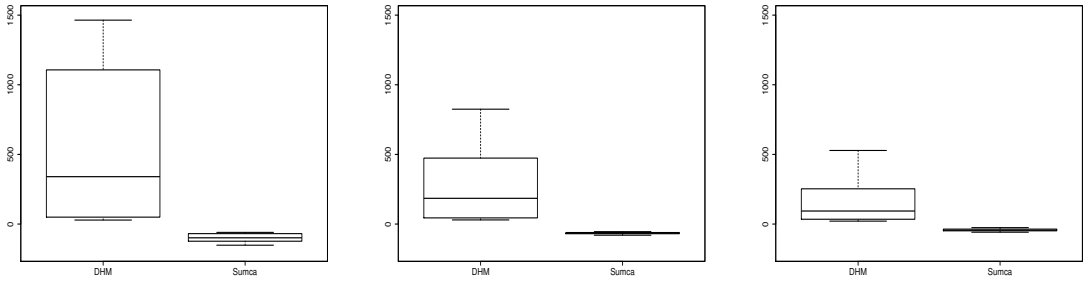
$m = 20$ $m = 50$ $m = 100$

$A = 0$



$m = 20$ $m = 50$ $m = 100$

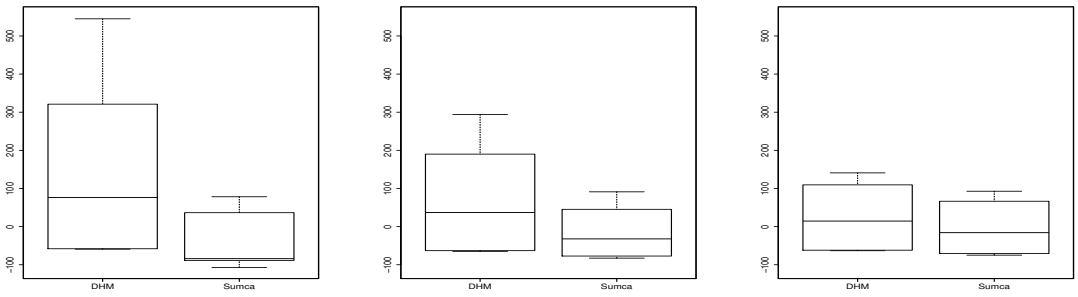
$A = 0.5$



$m = 20$ $m = 50$ $m = 100$

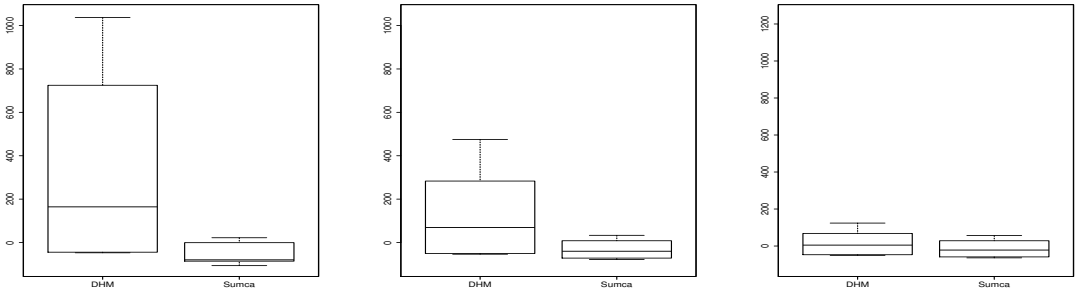
$A = 1$

Figure 2: Boxplots of % RB for Sumca and DHM methods; no model misspecification



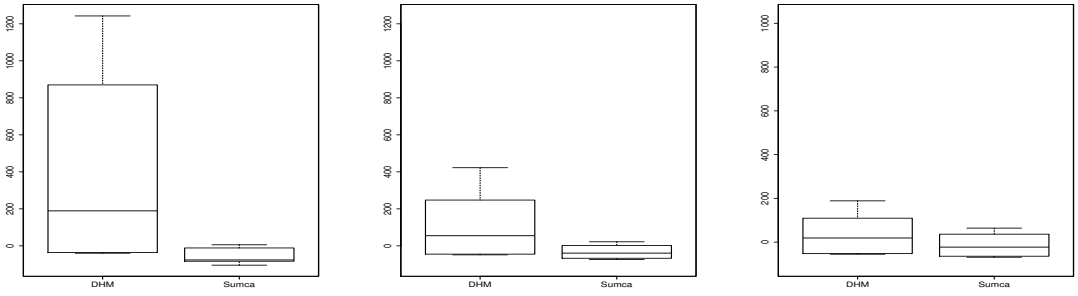
$m = 20$ $m = 50$ $m = 100$

$A = 0$



$m = 20$ $m = 50$ $m = 100$

$A = 0.5$



$m = 20$ $m = 50$ $m = 100$

$A = 1$

Figure 3: Boxplots of % RB for Sumca and DHM methods; model misspecification

The procedure is computationally very intensive. Jiang *et al.* (2018) proposed the McJack method, which requires the Monte-Carlo sample size, K, to satisfy $m^2/K \to 0$. However, for the Sumca estimator it is only suggested that $K = m$ (see the next paragraph). The computational saving by Sumca is mainly due to the following simple fact: If a term is already $O(m^{-1})$, a simple plug-in estimator will "do the trick", that is, produce an estimator of the term whose bias is $o(m^{-1})$; in other words, no bias correction is needed for this term. For Sumca, this term is $d(\psi) = b(\psi) - c(\psi)$ defined above (4), and no bias correction is needed after the plugging-in [see (4)]. The more intensive computational efforts of DB and McJack are spent on bias-correcting terms of $O(m^{-1})$, which is not necessary, so far as the second-order unbiasedness is concerned. In Sumca, the "plugging-in principle" is executed with the assistance of Monte-Carlo simulation.

As mentioned (second paragraph of Section 1), the current SAE literature has focused on (approximate) unbiasedness in MSPE estimation; little has been done regarding variance in the MSPE estimation. It was noted that the choice of K does not affect the second-order unbiasedness of the Sumca estimator. The following discussion is regarding the variance aspect. First, the Monte-Carlo part of the Sumca estimator, that is, the second term on the right side of (13), is not the main contributor to the variance. To see this, note that, with respect to the joint distribution of the data and Monte-Carlo sampling, we have

$$\mathrm{var}(\widehat{\mathrm{MSPE}}_K) = \mathrm{var}\left\{\mathrm{E}(\widehat{\mathrm{MSPE}}_K|y)\right\} + \mathrm{E}\left\{\mathrm{var}(\widehat{\mathrm{MSPE}}_K|y)\right\}. \tag{33}$$

Furthermore, by the way that the Monte-Carlo samples are drawn, we have

$$\mathrm{var}(\widehat{\mathrm{MSPE}}_K|y) \;\; = \;\; K^{-1}\mathrm{var}\{a(y_{[1]},\hat{\psi}) - a(y_{[1]},\hat{\psi}_{[1]})|y\}, \tag{34}$$

hence, the second term on the right side of (33) is $O(K^{-1})$. On the other hand, we have

$$\mathrm{E}(\widehat{\mathrm{MSPE}}_K|y) = a(y,\hat{\psi}) + \mathrm{E}\{a(y_{[1]},\hat{\psi}) - a(y_{[1]},\hat{\psi}_{[1]})|y\} = a(y,\hat{\psi}) + d(\hat{\psi}) \tag{35}$$

by the definition of $d(\psi)$ and the way the Monte-Carlo samples are drawn. The variance of $d(\hat{\psi})$ is typically $O(m^{-1})$. Thus, by (33)–(35), we have, by the Cauchy-Schwarz inequality

$$\mathrm{var}(\widehat{\mathrm{MSPE}}_K) = \mathrm{var}\{a(y,\hat{\psi})\} + O(m^{-1/2})\sqrt{\mathrm{var}\{a(y,\hat{\psi})\}} + O(m^{-1} + K^{-1}). \quad (36)$$

According to our earlier result [see the paragraph between (7) and (11)], the order of $\mathrm{var}\{a(y,\hat{\psi})\}$ is $O(m^{-1})$ under the general linear mixed model (10), with θ being a linear mixed effect, and $\hat{\theta}$ the EBP of θ. In a PMS situation, suppose that $\hat{\theta}$ is the EBP under the selected model, $\hat{M}$. If the model selection procedure is consistent, then with high probability (e.g., Theorem 5 of Jiang and Torabi 2019), one has $\hat{M} = M_{\mathrm{o}}$, where M_{o} the true underlying model. It follows that, with high probability, $\hat{\theta}$ becomes the EBP; therefore, the order $O(m^{-1})$ is still expected for $\mathrm{var}\{a(y,\hat{\psi})\}$ in this situation.

In general, as long as $\mathrm{var}\{a(y,\hat{\psi})\} = O(m^{-1})$ holds, by (36), to ensure that the order of the variance of the Sumca estimator does not increase due to K, the latter should be of the same order as m. A simple choice would be $K = m$.

On the other hand, it should be noted that $\mathrm{var}\{a(y,\hat{\psi})\} = O(m^{-1})$ is not always expected to hold. For example, in the case of mixed logistic model, discussed in Section 4.3, one has the expression $a(y,\hat{\psi}) = \mathrm{var}(\theta_i|y)|_{\psi=\hat{\psi}}$, which depends on $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$, in addition to $\hat{\psi}$. Thus, in this case, the variance of $a(y,\hat{\psi})$ is expected to be $O((m \wedge n_i)^{-1})$ rather than $O(m^{-1})$. However, by (36), one can see that $K = m$ is still a good choice to ensure that the variance contribution due to the Monte-Carlo part is, at most, the same order as that due to the leading term, $a(y,\hat{\psi})$.

We shall explore the variance issue more broadly, including the situation where the variance of $a(y,\hat{\psi})$ may not be $O(m^{-1})$, with a rigorous treatment, in our future work.

Another issue regarding the MSPE estimation is how to obtain accurate MSPE estimation under model misspecification. In the case of PMS (see Section 3), our approach is based on existence of a full model, which is a correct model. However, the issue can also

arise in a non-model-selection context, where the underlying model is misspecified. See Liu, Ma and Jiang (2019).

Acknowledgements. The research of Jiming Jiang is partially supported by the NSF grant DMS-1510219. The research of Mahmoud Torabi is supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors are grateful to an Associate Editor and two referees for their constructive comments and suggestions that have led to major improvement of the work as well as presentation.

References

[1] Das, K. and Jiang, J. and Rao, J. N. K. (2004), Mean squared error of empirical predictor, *Ann. Statist.* 32, 818-840.

[2] Datta, G. S., Hall, P., and Mandal, A. (2011), Model selection by testing for the presence of small-area effects, and applications to area-level data, *J. Amer. Statist. Assoc.* 106, 361-374.

[3] Datta, G. S. and Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica* 10, 613-627.

[4] Datta, G. S. and Rao, J. N. K. and Smith, D. D. (2005), On measuring the variability of small area estimators under a basic area level model, *Biometrika* 92, 183-196.

[5] Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: an application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.* 74, 269-277.

[6] Hall, P. and Maiti, T. (2006a), Nonparametric estimation of mean-squared prediction error in nested-error regression models, *Ann. Stat.* 34, 1733-1750.

[7] Hall, P. and Maiti, T. (2006b), On parametric bootstrap methods for small area prediction, *J. Roy. Statist. Soc. Ser. B* 68, 221-238.

[8] Jiang, J. (2001), Mixed effects models with random cluster sizes, *Statist. Probab. Letters* 53, 201–206.

[9] Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

[10] Jiang, J. (2010), *Large Sample Techniques for Statistics*, Springer, New York.

[11] Jiang, J. (2017), *Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems*, Chapman & Hall/CRC.

[12] Jiang, J. and Lahiri, P. (2001), Empirical best prediction for small area inference with binary data, *Ann. Inst. Statist. Math.* 53, 217-243.

[13] Jiang, J., Lahiri, P. and Nguyen, T. (2018), A unified Monte-Carlo jackknife for small area estimation after model selection, *Ann. Math. Sci. Appl.* 3, 405-438.

[14] Jiang, J., Lahiri, P. and Wan, S. (2002), A unified jackknife theory for empirical best prediction with M-estimation, *Ann. Statist.* 30, 1782-1810.

[15] Jiang, J., Nguyen, T., and Rao, J. S. (2015), The E-MS algorithm: Model selection with incomplete data, *J. Amer. Statist. Assoc.* 110, 1136-1147.

[16] Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008), Fence methods for mixed model selection, *Ann. Statist.* 36, 1669-1692.

[17] Jiang, J. and Torabi (2019), Sumca: Simple, unified, Monte-Carlo assisted approach to second-order unbiased MSPE estimation, Technical Report.

[18] Liu, X., Ma, H. and Jiang, J. (2019), Assessing uncertainty involving model misspecification: A one-bring-one route, Technical Report.

[19] Molina, I., Rao, J.N.K. and Datta, G.S. (2015), Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects, *Surv. Methodol.* 41, 1-19.

[20] Müller, S., Scealy, J. L., and Welsh, A. H. (2013), Model selection in linear mixed models, *Statist. Sci.* 28, 135-167.

[21] Pfeffermann, D. (2013), New important developments in small area estimation, *Statist. Sci.* 28, 40-68.

[22] Prasad, N. G. N. and Rao, J. N. K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 85, 163-171.

[23] Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation*, 2nd ed., Wiley, New York.

[24] Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer, New York.

[25] Slud, E.V. and Maiti, T. (2006), Mean-squared error estimation in transformed Fay-Herriot models, *J. Roy. Statist. Soc. Ser. B* 68, 239-257.

[26] Tibshirani, R. J. (1996), Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B* 58, 267-288.